An Automatic Grading System for Neonatal Endotracheal Intubation with Multi-Task Convolutional Neural Network

Yan Meng[1] and James K. Hahn[2]

***Abstract*— Neonatal endotracheal intubation (ETI) is an intricate medical procedure that poses considerable challenges, demanding comprehensive training to effectively address potential complications in clinical practice. However, due to limited access to clinical opportunities, ETI training relies heavily on physical manikins to develop a certain level of competence before clinical exposure. Nonetheless, traditional training methods prove ineffective due to scarcity of expert instructors and the absence of internal situational awareness within the manikins, preventing thorough performance assessment for both trainees and instructors. To address this gap, there is a need to develop an automatic grading system that can assist trainees in performance assessment. In this paper, we proposed a multi-task Convolutional Neural Network (MTCNN) based model for assessing ETI proficiency, specifically targeting key performance features recommended by expert instructors. The model comprises three modules: an ETI simulation module that captures the ETI procedures performed on a standard neonatal task trainer manikin, an automatic grading module that extracts and grades the identified key performance features, and a data visualization module that presents the assessment results in a user-friendly manner. The experimental results demonstrated that the proposed automatic grading system achieved an average classification accuracy of 93.6%. This study established the successful integration of intuitive observed features with latent features derived from multivariate time series (MTS) data, coupled with multi-task deep learning techniques, for the automatic assessment of ETI performance.**

***Clinical relevance*— The proposed automatic grading system facilitates an enhanced neonatal endotracheal intubation training experience for neonatologists.**

I. INTRODUCTION

Neonatal endotracheal intubation (ETI) is a widely employed medical procedure in newborn infants aimed at establishing an artificial airway. This procedure becomes necessary when newborns are incapable of achieving sufficient respiration autonomously or when mechanical ventilation assistance is required. The intricacy of neonatal ETI arises from the narrow and delicate nature of the airways, the limited space within the oral cavity, and the imperative need for prompt action. The success of neonatal ETI is evaluated based on the operator's proficiency in correctly placing the endotracheal tube within a time frame of 30 seconds [1]. In order to attain the requisite level of competence prior to clinical exposure, conventional approaches have relied on task trainer manikins in conjunction with instructor feedback. Nevertheless, these methods have inherent limitations, including restricted visibility of the neonatal airway, large class sizes, and time constraints during training sessions, thereby hindering instructors' ability to identify specific causes of procedural failure for targeted feedback. As a result, there is a rising demand for an automated training framework that can rapidly identify deficiencies in learners and enhance the learning process by providing intuitive feedback [2].

In the field of medical training, machine learning (ML)[3], [4] and deep learning (DL)[5], [6] techniques have gained considerable popularity for their application in making diagnostic decisions and assessing performance in medical procedures. Traditional machine learning algorithms typically analyze features that are designed based on the expertise of medical professionals. However, the extraction of features in machine learning methods is typically limited to specific data points, such as the maximum force, optimal rotation angle, and completion time of a procedure, which may lead to a loss of comprehensive information concerning the entire medical procedure. In contrast, deep learning methods automatically extract high-dimensional features pertaining to the entire procedure, resulting in improved learning outcomes. Nevertheless, the lack of interpretability associated with deep learning hampers its clinical application. The fusion of features derived from both machine learning and deep learning methods enables the exploitation of the distinct strengths inherent in each approach, leading to a substantial enhancement in the overall performance of medical training.

For ETI training, the automatic grading of key motion features serves as a valuable and intuitive form of feedback, aiding trainees to identify areas for improvement [7]. However, traditional machine learning methods, such as the multilayer perceptron (MLP), have inherent limitations as they rely solely on the values of extracted features and their associated labels for classification. Whereas contemporary techniques, such as multi-task learning with neural networks [8], [9], offer a superior alternative by leveraging motion information from the entire multivariate time series (MTS) dataset, and facilitating knowledge sharing among different tasks. In this paper, we proposed a multi-task convolutional neural network (MTCNN) structure to facilitate the assessment of ETI performances and the grading of motion features on a 5-point scale. By jointly training on multiple tasks, the deep learning networks demonstrated the advantage of effectively capturing underlying patterns, dependencies, and relationships across diverse tasks, which, in turn, enabled a substantial improvement in the model's performance and its generalization capabilities.

[1]Yan Meng is with the Department of Computer Science, The George Washington University, Washington, DC 20052, USA mengy@gwu.edu

[2]James K. Hahn is with the Department of Computer Science, The George Washington University, Washington, DC 20052, USA hahn@gwu.edu

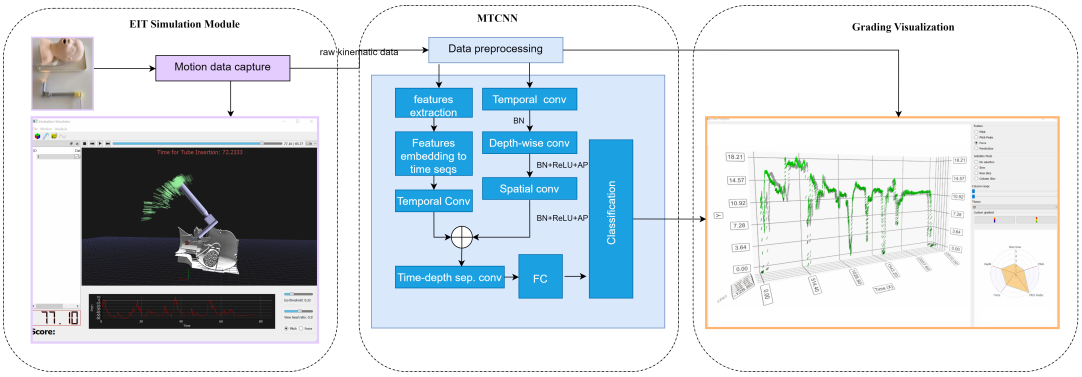


Fig. 1: The overview of the automatic grading system for neonatal endotracheal intubation.

II. METHODS

The proposed automatic grading system for ETI training consists of three modules: the ETI simulation module, the automatic grading framework MTCNN, and the grading feedback visualization module. An overview of the system architecture is given in Fig. 1.

A. Dataset Collection

For the purpose of this study, a dataset comprising 193 trials conducted by 45 subjects was collected. The motions of both the standard full-term Laerdal® task trainer manikin and the laryngoscope were precisely tracked using TrakSTAR™ EM sensors, and both the manikin and the laryngoscope were registered to their CT scanned virtual avatars, respectively.

Approval for the study was obtained from the Institutional Review Board of the Children's National Hospital. The subjects enrolled in the study comprised pediatric residents and neonatologists with diverse levels of expertise, and each subject performed 3 to 5 trials. Following the completion of the trials, the 3D motions were randomly played back and evaluated by three expert graders with extensive experience of over 6 years as practicing neonatologists.

B. Data Preprocessing and Feature Selection

The 3D kinematic MTS data were temporally aligned to have a uniform length T. In order to eliminate the variation in motions across different trials of the manikin head, the relative transformations between the laryngoscope and the head were calculated. Each individual trial $S = \{x_1, x_2, ..., x_T\}$ encompassed the position and orientation of the laryngoscope at every sampled time step, where the rotation was represented as quaternion. The two features were concatenated into a motion vector x with 7 dimensions as the input to the MTCNN.

The key performance features selected by experienced instructors encompassed several factors: procedure duration, number of attempts, duration the glottis is visible, laryngoscope penetration depth into the tongue plane, force applied to the gum, rotation peaks, and rotation rocking. Where rotation with respect to each axis are pitch, yaw, and roll respectively. Previous study [10] investigating the significance of these features in relation to the final ETI performance indicated that the peaks and rocking of yaw and roll had negligible impact on the overall scores of the procedure. Furthermore, the features of duration and attempts are deterministic upon the completion of the procedure. A trial that exceeds 30 seconds or involves more than one attempt is considered a failed trial, eliminating the need for machine learning algorithms to grade these features. Therefore, to evaluate the smoothness and stability of an ETI procedure, we focused solely on grading five key motion features: duration the glottis is visible, penetration depth, force, pitch peak, and pitch rocking. In order to integrate distinct feature types derived from both ML and DL approaches, an effective method was developed to aggregate these features. We constructed a new time sequence $G = \{g_1, g_2, ..., g_T\}$ and inserted the key feature values at their corresponding time steps within the original MTS data S, then fed this feature sequence to the MTCNN.

C. Multi-Task Convolutional Neural Network (MTCNN)

MTCNN leveraged the latent features extracted from the motion kinematic data to rate the key performance features that contribute to the success of an ETI procedure. The network received inputs S and G, and generated output values representing the grades associated with each key performance feature. The architecture of the proposed network is depicted in Fig. 2. The input and output of each layer, excluding the fully connected layers, were organized into 3D data blocks with dimensions $\{L, C, F\}$, where L and C represent the spatial width and height of the feature map, respectively, and F represents the number of feature maps.

1) Latent feature extraction: The 3D kinematic data were mapped into a higher-level feature space using a series of three convolutional layers: temporal, depth-wise, and spatial convolution. In the first layer, temporal filtering was performed with F_1 kernels of size $(1, k_1)$. The depth-wise convolution layer captured the relative information among different input channels. Subsequently, the spatial convolution layer further extracted information across different feature maps using F_2 kernels of size $(1, k_2)$, where $F_1 = 16, F_2 = 32$, and all ks were set to 7 in our study. Following each convolutional layer, batch normalization and average pooling were applied to standardize the features, reduce their

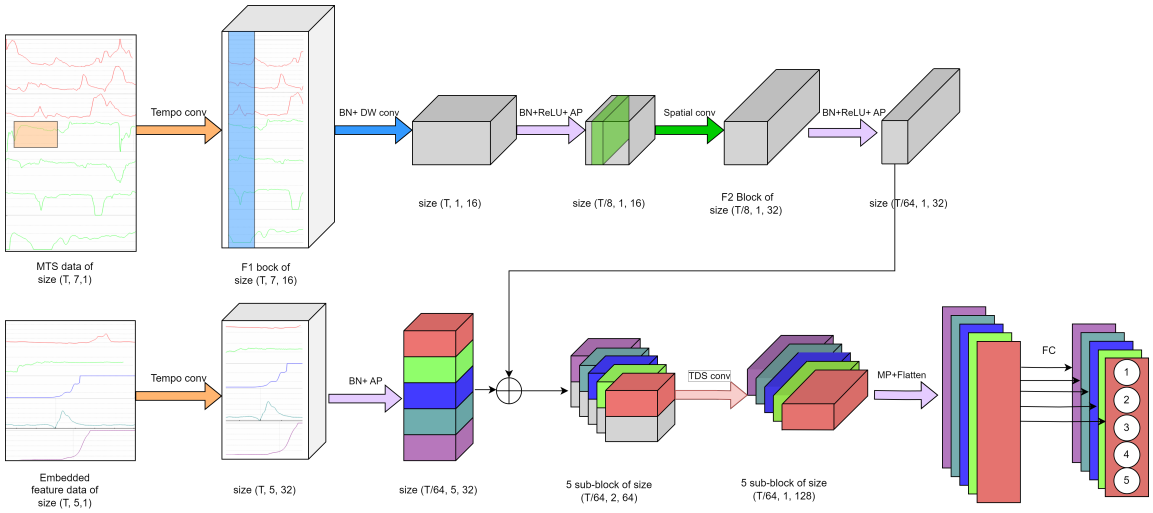


Fig. 2: MTCNN architecture. The abbreviations BN, DW, conv, AP, MP, and TDS represent batch normalization, depth-wise, convolution, average pooling, max pooling, and time-depth separable respectively.

size, and accelerate network training.

2) Feature fusion: To ensure data compatibility, the embedded feature MTS data were mapped to a feature space through a temporal convolution layer, batch normalization and average pooling. The output features maps for each key performance feature were concatenated with the latent features separately, resulting in fused feature maps that can be expressed as:

$$v = \alpha v_k \oplus (1-\alpha) v_l \tag{1}$$

Where v_k is the key performance feature, v_l is the latent feature, and $\alpha \in (0, 1)$, is a weight parameter that adjusts the contribution of the key performance feature. In particular, when $\alpha = 0$, the final grading classification relies solely on the kinematic data, while $\alpha = 1$ indicates that the network exhibits behavior similar to that of a multilayer perceptron.

3) Classification: Each group of fused feature maps were fed into two time-depth separable (TDS) convolutional blocks [11]. The TDS model has the advantages of high training efficiency, wide receptive field, and enhanced classification accuracy compared to prevalent attention-based strategies. Finally, the output of TDS block were directed to a fully connected layer with a softmax activation function to generate grade prediction for each key performance feature.

4) Training: The five tasks were jointly trained with cross-entropy loss as the optimization objective:

$$E(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{M} \sum_{j=1}^{N} y_i \log \hat{y} \tag{2}$$

Where M is the number of tasks, N is the number of classes, y is the classification target class defined in the labeling phase, and $\hat{y}$ is the predicted class.

The networks were trained and tested with a NVIDIA GTX 1080Ti GPU. We used the Adam optimizer with a learning rate of 0.005, a batch size of 16, and a categorical cross-entropy loss over 500 epochs. The dropout rate for TDS blocks was 0.2, and a label smoothing factor of 0.3 was applied to prevent overfitting.

D. Grading Feedback Visualization

The grading feedback visualization module presented the classification results in a polar chart, accompanied by key performance feature values over time. The polar chart was constructed using a 5-point scale, where each radial axis corresponded to a specific feature. The area enclosed by the grade assigned to each feature served as an indicator of the trainee's overall performance, with a larger area signifying a higher level of proficiency exhibited by the trainee. An overview of the visualization is depicted in Fig. 3. This visual representation offered trainees an intuitive feedback mechanism and provided guidance on areas in need of improvement.

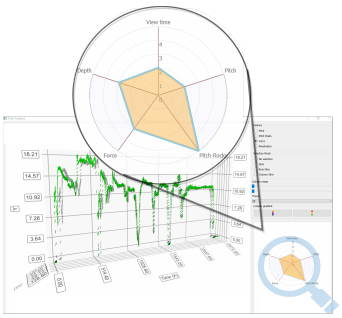


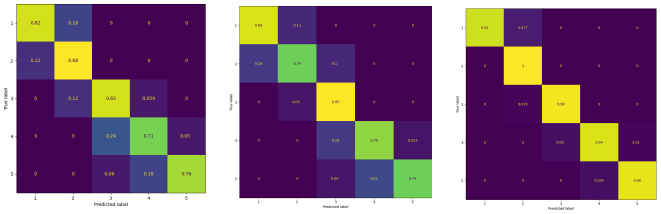
Fig. 3: The grading feedback visualization module, where the classification results are presented in a polar chart.

III. RESULTS

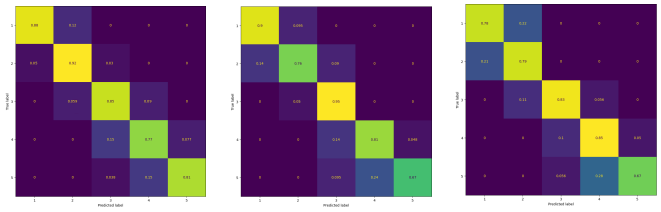
The MTCNN performance was assessed using accuracy and F-score metrics, and the results were presented in Table I. The results were the averaged classification output over one round of the leave-one-out cross validation. We examined the

TABLE I: Classification results comparison

Metrics	MTCNN with varying α					MLP
	$\alpha = 0$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 1$	
Accuracy(%)	85.1±1.8	89.3±1.5	93.6±1.06	90.9±1.2	88.4±1.3	84.2±2.1
F-score	0.76±0.04	0.79±0.02	0.86±0.03	0.81±0.02	0.83±0.02	0.75±0.3



(a) $\alpha = 0$ (b) $\alpha = 0.3$ (c) $\alpha = 0.5$



(d) $\alpha = 0.7$ (e) $\alpha = 1$ (f) Baseline MLP

Fig. 4: Confusion Matrices for the force feature classification.

impact of weight α on classification results, and compared the proposed method with a baseline MLP.

The numerical analysis revealed that the performance of MTCNN improved as the weight α of the key performance features increased, which was a comprehensible outcome: the embedded key performance MTS data were equivalent to prior information that effectively enriched the network's knowledge base. The network reached its peak performance when both the kinematic data and embedded feature data contributed equally to the classification task. The average classification accuracy reached 93.6%, surpassing the MTCNN with only kinematic data input by 10% and outperforming the baseline MLP by 12%.

However, as α continued to increase, the network gradually lost kinematic motion information, resulting in a decrease in accuracy. Nevertheless, the MTCNN with only embedded key performance features still performed slightly better than the MTCNN with only kinematic data as input. This was attributed to the embedded key performance feature data containing direct observations related to the classification target. Overall, the performance of MTCNN with varying α values surpassed that of the traditional MLP method. Our conclusion was further supported by the sampled confusion matrices depicted in Figure 4, which reflected the force feature classification performance of different network settings and the baseline MLP. Notably, the MTCNN outperformed the baseline MLP across all classes, except for class 4, where the MLP exhibited slightly higher accuracy than the MTCNN with $\alpha = 0$.

IV. CONCLUSIONS

In this paper, we presented a multi-task convolutional neural network for rating the key performance features of ETI procedures. Our network offered the advantage of fusing both kinematic motion data and predefined features with critical contextual prior knowledge to train an automatic grader, thereby enabling the provision of interpretable feedback to trainees. Through jointly training on multi-tasks, the network learned to balance the shared knowledge and focus on the common underlying structure, thus preventing overfitting and improving generalization on small dataset. The experimental results demonstrated that the MTCNN achieved consistent and high accuracy in the classification tasks. In the future, we aim to explore additional intuitive feedback strategies that provide guidance for improvement to trainees, such as real-time motion guidance and textual descriptor motion cues, and to perform user studies.

ACKNOWLEDGMENT

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number R01HD091179. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

[1] *Textbook of Neonatal Resuscitation, 8th Ed.* American Academy of Pediatrics and American Heart Association, 2021.

[2] Y. Meng, "Vr enhanced interactive intubation simulation," in *Proceedings of the International Conference on Modeling, Simulation and Visualization Methods (MSV)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2018, pp. 81–87.

[3] M. J. Fard, S. Ameri, R. B. Chinnam, A. K. Pandya, M. D. Klein, and R. D. Ellis, "Machine learning approach for skill evaluation in robotic-assisted surgery," *arXiv preprint arXiv:1611.05136*, 2016.

[4] A. Ghasemloonia, Y. Maddahi, K. Zareinia, S. Lama, J. C. Dort, and G. R. Sutherland, "Surgical skill assessment using motion quality and smoothness," *Journal of surgical education*, vol. 74, no. 2, pp. 295–305, 2017.

[5] X. A. Nguyen, D. Ljuhar, M. Pacilli, R. M. Nataraja, and S. Chauhan, "Surgical skill levels: Classification and analysis using deep neural network model and motion signals," *Computer methods and programs in biomedicine*, vol. 177, pp. 1–8, 2019.

[6] Z. Wang and A. Majewicz Fey, "Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery," *International journal of computer assisted radiology and surgery*, vol. 13, pp. 1959–1970, 2018.

[7] Y. Meng, "Force aware haptic rendering for intubation simulation," *International Journal of Engineering & Technology*, vol. 7, no. 3, pp. 61–66, 2018.

[8] K. Crammer and Y. Mansour, "Learning multiple tasks using shared hypotheses," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[9] V. Andrearczyk, P. Fontaine, V. Oreiller, J. Castelli, M. Jreige, J. O. Prior, and A. Depeursinge, "Multi-task deep segmentation and radiomics for automatic prognosis in head and neck cancer," in *Predictive Intelligence in Medicine: 4th International Workshop, PRIME 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 4*. Springer, 2021, pp. 147–156.

[10] S. Zhao, W. Li, X. Zhang, X. Xiao, Y. Meng, J. Philbeck, N. Younes, R. Alahmadi, L. Soghier, and J. Hahn, "Automated assessment system with cross reality for neonatal endotracheal intubation training," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2020, pp. 738–739.

[11] A. Hannun, A. Lee, Q. Xu, and R. Collobert, "Sequence-to-sequence speech recognition with time-depth separable convolutions," *arXiv preprint arXiv:1904.02619*, 2019.